

Vorwort	IX
Einleitung	XI
1 Schnelleinstieg in Solr	1
Was ist Solr?	1
Was ist Lucene?	2
Was kann Solr?	2
Features	2
Community	5
Einsatzmöglichkeiten	5
Ein kurzer historischer Rückblick	6
Also ist Solr Google?	6
Erste Schritte – Solr entpacken und starten	7
Apache Solr herunterladen	7
Apache Solr starten	7
Inhalte indexieren	10
Die erste Suche	11
Die ersten Ergebnisse	13
2 Die Konzepte von Solr kennenlernen	15
Apache Solr innerhalb einer Applikation	15
Der Grundaufbau von Solr	17
Indexierung – UpdateRequestHandler	17
Suche und Suchfeatures – SearchHandler	18
Echtzeitsuche – RealTimeGetHandler	21
Rückgabe des Ergebnisses – ResponseWriter	23

Administration – AdminHandler	25
Ausfallsicherheit – ReplicationHandler	35
Architektur innerhalb einer Solr-Instanz – CoreAdminHandler	35
Die Indexierung – Out-of-the-Box-Möglichkeiten	45
Indexierung von XML-, CSV-, JSON-Dateien – UpdateRequestHandler . . .	45
Extraktion von Inhalt – Solr Cell – ExtractingRequestHandler	48
Datenbanken, RSS, Wikipedia, ... indexieren – DataImportHandler	50
Beeinflussung des Indexierungsprozesses – UpdateRequestProcessors	59
Die Suche – Wie kann ich suchen?	67
Einfache Suche – Termsuche	67
Boolesche Operatoren	68
Einsatz von Wildcards	70
Query Parser	71
Fuzzy Search – Ungenaue Suche	80
Phrasensuche mit dem Lucene Query Parser	81
Range-Queries	81
Filter-Queries	82
Sortierung	82
3 Den Index konfigurieren	85
Der Lucene-Index	85
Operationen auf dem Index	86
Die Schema-Konfiguration	90
Solr-Feldtypen	91
Solr-Felder	94
Allgemeine Einstellungen	96
Der Analyse-Prozess	98
Konfiguration des Analyse-Prozesses	98
1. Schritt: CharFilter	100
2. Schritt: Tokenizer	102
3. Schritt: TokenFilter	106
Typische Anwendungsfälle der Analyse	117
Die sprachspezifische Analyse	121
Das Analyse-Interface	124
Ein Blick in den Index	128
4 Was kann Solr out-of-the-box?	133
Die Konfigurationsdatei solrconfig.xml	133
Allgemeine Einstellungen	134

Index-Einstellungen	135
Query-Einstellungen	136
Request Dispatcher-Einstellungen	137
Konfiguration der Admin-Oberfläche	137
Such-Features out-of-the-box.	138
Velocity und der /browse-RequestHandler	138
Facetten – Suchergebnisse verfeinern	141
AutoSuggest – Suchbegriffe vorschlagen	151
Highlighting – Suchbegriffe im Treffer hervorheben	156
Result Grouping – ähnliche Dokumente gruppieren	160
Meinten Sie ... – Tippfehler ausbessern	164
MoreLikeThis – ähnliche Dokumente finden.	169
Elevate – Top-Treffer definieren.	174
Terms-Komponente – Solr-Felder auslesen	178
TermVector-Komponente – Term-Informationen auswerten	183
Stats-Komponente – statistische Auswertung	186
/browse-RequestHandler für die Wikipedia.	188

5 Scoring und Relevanz beeinflussen 191

Precision versus Recall	191
Den Scoring-Mechanismus verstehen	193
Konstantes Scoring	193
Lucene-Scoring	194
Der TF-IDF-Algorithmus	194
Custom-TF-IDF-Scoring	197
Scoring-Probleme analysieren	198
Lucenes Explain TF-IDF-Funktionalität.	198
Das Scoring beeinflussen	203
Query Parser für die Scoring-Manipulation nutzen	203
Mit FunctionQueries das Scoring beeinflussen	209
Typische Scoring-Anwendungsfälle	215

6 Skalierung der Suche – die Solr-Architektur gestalten 219

Master/Slave-Architektur	219
Indexierung.	224
Replication	224
Mit Replication Backups erstellen	232
SolrCloud	234
Reichen Replication und Sharding nicht aus?	234

SolrCloud – Diese Gedanken stecken hinter dieser Innovation	234
Who is Who oder: Die Terminologie der SolrCloud	235
Zero-Installation – einfach loslegen	236
Indexieren und Suchen in der SolrCloud	239
Erhöhung der Ausfallsicherheit der Administrationsseite – externes ZooKeeper-Ensemble	240
Mehr Infos – clusterstate.json.	250
SolrCloud-Verwaltung – Collections-API.	252
Wohin gehen meine Dokumente? – Document Routing	263
Verwaltung mehrerer Collections in der SolrCloud	266
Pitfalls – Auf was Sie sonst noch achten sollten	268

7 Ein Blick über den Tellerrand 271

Mit Solr arbeiten – Client-APIs	271
Liste der verfügbaren Client-APIs.	272
Der Java-Client – SolrJ	273
Deployment von Solr in Apache Tomcat	274
Tomcat-Download	275
Tomcat-Installation	275
Solr-Deployment.	276
Monitoring Ihrer Solr-Installation	278
JMX-Aktivierung und Tools	278
Log-Auswertung	283
Die Community – Wie kann ich zum Projekt beitragen?	285
Die Apache Software Foundation	286
Apache Hadoop – Lösung für verteilte Systeme	287
Apache Mahout – Clustering, Klassifikation & Recommendations.	288
Apache Stanbol – Content Enrichment	291
Apache OpenNLP – Verarbeitung natürliche Sprache	293
Apache Nutch – Webseiten crawlen.	294
Apache ManifoldCF – flexibles Crawling-Framework	294
Die Konkurrenz – Elasticsearch	295
Die Deutsche Wikipedia mit Elasticsearch indexieren	295

Glossar 303

Index 313